(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Publication number: 0 535 807 A2

(12) EUROPEAN PATENT APPLICATION

(21) Application number: 92308079.0

(51) Int. Cl.5: G06F 15/16

(22) Date of filing: 07.09.92

(30) Priority: 01.10.91 US 769538

(43) Date of publication of application:
07.04.93 Bulletin 93/14

(84) Designated Contracting States:
DE FR GB IT

(71) Applicant: TANDEM COMPUTERS INCORPORATED
10435 N. Tantau Avenue
Cupertino, California 95014-0709(US)

(72) Inventor: Walker, Mark
20000 Gist Road
Los Gatos, California 95030(US)
Inventor: Lui, Albert S.
3164 Heritage Valley Drive
San Jose, California 95148(US)
Inventor: Sammer, Harald
Am Eschenhorst 10
W-6382 Friedrichsdorf(DE)
Inventor: Chan, Wing M.
7922 Kemper Court
Pleasanton, California 94588(US)
Inventor: Fuller, William T.
1536 Shasta Avenue
San Jose, California 95126(US)

(74) Representative: Allman, Peter John et al
MARKS & CLERK Suite 301 Sunlight House
Quay Street
Manchester M3 3JY (GB)

(54) Linear and orthogonal expansion of array storage in multiprocessor computing systems.

(57) A multiprocessing computer system with data storage array systems allowing for linear and orthogonal expansion of data storage capacity and bandwidth by means of a switching network coupled between the data storage array systems and the multiple processors. The switching network provides the ability for any CPU to be directly coupled to any data storage array. By using the switching network to couple multiple CPU's to multiple data storage array systems, the computer system can be configured to optimally match the I/O bandwidth of the data storage array systems to the I/O performance of the CPU's.

EP 0 535 807 A2

BACKGROUND OF THE INVENTION

1. *Field of the Invention*

This invention relates to computer systems, and more particularly to a multiprocessing computer system with data storage array systems allowing for linear and orthogonal expansion of data storage capacity and bandwidth by means of a switching network.

2. *Description of Related Art*

A typical multiprocessing computer system generally involves one or more data storage units which are connected to a plurality of Central Processor Units (CPU's), either directly through an input/output (I/O) bus, or through an I/O control unit and one or more I/O channels. The function of the data storage units is to store data and programs which the CPU's use in performing particular data processing tasks.

One type of multiprocessing system known in the art is described in U.S. Patent No. 4,228,496, assigned to the assignee of the present invention. A simplified version of the computer system architecture taught in that patent is shown in FIGURE 1. The system shown therein provides for a high degree of reliability by providing two redundant interprocessor busses IPB interconnecting a plurality of CPU's 1. However, where cost predominates over reliability concerns, a single interprocessor bus may be used in a multiprocessing system.

The system shown in FIGURE 1 includes a plurality of data storage units 2 each coupled to at least two CPU's 1 by means of an I/O bus 3 (or, alternatively, through redundant I/O control units). Various type of data storage units are used in such a data processing system. A typical system may include one or more large capacity tape units and/or disk drives (magnetic, optical, or semiconductor). Again, if cost is a predominant factor, single connections rather than dual connections can be used.

Any CPU 1 in the architecture can access any directly coupled data storage unit 2. In addition, any CPU 1 in the architecture can access any other data storage unit 2 indirectly over the IPB via another CPU 1.

The architecture shown in FIGURE 1 allows for linear expansion of computing resources by adding CPU's 1 to the interprocessor bus IPB, in the "x" direction (see FIGURE 1). The architecture also allows for linear expansion of I/O resources by adding data storage units 2 to the I/O busses or channels, in the orthogonal "y" direction. Expansion in the x and y directions can be independent of each other, limited only by performance and physical constraints.

Thus, the current art provides for linear expansion of CPU's and orthogonal and linear expansion of individual data storage units 2 to correspond to the storage requirements of the CPU's.

More recently, highly reliable disk array data storage systems have been introduced to the market. Such disk array systems present a challenge when coupled within such a multiprocessor architecture.

Disk array systems are of various types. A research group at the University of California, Berkeley, in a paper entitled "A Case for Redundant Arrays of Inexpensive Disks (RAID)", Patterson, *et al.*, *Proc. ACM SIGMOD*, June 1988, has catalogued a number of different types by defining five architectures under the acronym "RAID" (for Redundant Arrays of Inexpensive Disks).

A RAID 1 architecture involves providing a duplicate set of "mirror" data storage units and keeping a duplicate copy of all data on each pair of data storage units. A number of implementations of RAID 1 architectures have been made, in particular by Tandem Computers Incorporated.

A RAID 2 architecture stores each bit of each word of data, plus Error Detection and Correction (EDC) bits for each word, on separate disk drives. For example, U.S. Patent No. 4,722,085 to Flora *et al.* discloses a disk drive memory using a plurality of relatively small, independently operating disk subsystems to function as a large, high capacity disk drive having an unusually high fault tolerance and a very high data transfer bandwidth. A data organizer adds 7 EDC bits (determined using the well-known Hamming code) to each 32-bit data word to provide error detection and error correction capability. The resultant 39-bit word is written, one bit per disk drive, on to 39 disk drives. If one of the 39 disk drives fails, the remaining 38 bits of each stored 39-bit word can be used to reconstruct each 32-bit data word on a word-by-word basis as each data word is read from the disk drives, thereby obtaining fault tolerance.

A RAID 3 architecture is based on the concept that each disk drive storage unit has internal means for detecting a fault or data error. Therefore, it is not necessary to store extra information to detect the location of an error; a simpler form of parity-based error correction can thus be used. In this approach, the contents of all storage units subject to failure are "Exclusive OR'd" (XOR'd) to generate parity information. The resulting parity information is stored in a single redundant storage unit. If a storage unit fails, the data on that unit can be reconstructed on to a replacement storage unit by XOR'ing the data from the remaining storage units with the parity information. Such an arrangement has the advantage over the mirrored disk RAID 1

architecture in that only one additional storage unit is required for "N" storage units. A further aspect of the RAID 3 architecture is that the disk drives are operated in a coupled manner, similar to a RAID 2 system, and a single disk drive is designated as the parity unit. One implementation of a RAID 3 architecture is the Micropolis Corporation Parallel Drive Array, Model 1804 SCSI, which uses four parallel, synchronized disk drives and one redundant parity drive. The failure of one of the four data disk drives can be remedied by the use of the parity bits stored on the parity disk drive. Another example of a RAID 3 system is described in U.S. Patent No. 4,092,732 to Ouchi.

A RAID 4 architecture uses the same parity error correction concept of the RAID 3 architecture, but improves on the performance of a RAID 3 system with respect to random reading of small files by "uncoupling" the operation of the individual disk drive actuators, and reading and writing a larger minimum amount of data (typically, a disk sector) to each disk (this is also known as block striping). A further aspect of the RAID 4 architecture is that a single storage unit is designated as the parity unit.

A RAID 5 architecture uses the same parity error correction concept of the RAID 4 architecture and independent actuators, but improves on the writing performance of a RAID 4 system by distributing the data and parity information across all of the available disk drives. Typically, "N + 1" storage units in a set (also known as a "redundancy group") are divided into a plurality of equally sized address areas referred to as blocks. Each storage unit generally contains the same number of blocks. Blocks from each storage unit in a redundancy group having the same unit address ranges are referred to as "stripes". Each stripe has N blocks of data, plus one parity block on one storage unit containing parity for the remainder of the stripe. Further stripes each have a parity block, the parity blocks being distributed on different storage units. Parity updating activity associated with every modification of data in a redundancy group is therefore distributed over the different storage units. No single unit is burdened with all of the parity update activity. For example, in a RAID 5 system comprising 5 disk drives, the parity information for the first stripe of blocks may be written to the fifth drive; the parity information for the second stripe of blocks may be written to the fourth drive; the parity information for the third stripe of blocks may be written to the third drive; etc. The parity block for succeeding stripes typically "precesses" around the disk drives in a helical pattern (although other patterns may be used). Thus, no single disk drive is used for storing the parity information, as in the RAID 4 architecture. An example of a RAID 5 system is described in U.S. Patent No. 4,761,785 to Clark *et al.*

The challenge posed in coupling disk array data storage systems to a multiprocessor architecture that provides for linear and orthogonal CPU and data storage expansion is in matching the I/O bandwidth of the disk array system to the I/O capacity of the coupled CPU's. Because of the overhead cost of the array controller needed to manage a disk array, many data storage units are required within the array to achieve cost benefits by spreading the controller cost over multiple data storage units. Additionally, overall disk array system performance increases linearly with the number of data storage units within the system. Therefore, a typical disk array system includes an array controller and 3 or more disks (in some configurations, dozens of disks may be attached). However, the large number of disks in a typical disk array system often results in the array system having greater I/O performance (i.e., data transfers per second) than a single CPU can accommodate, leading to under-utilization of the data transfer capacity of the data storage units. As a consequence, the CPU's directly attached to a disk array system become a bottleneck for indirect accesses to the array from other CPU's. Adding additional disk array systems to other CPU's does not resolve the bottleneck problem with respect to data stored in a disk array system that is not directly coupled to such CPU's. Such an approach is also costly because the extra data transfer capacity of each disk array is not used.

It is thus difficult to match the I/O bandwidth of a disk array system to the I/O performance of multiple CPU's in a traditional multiprocessor computer system having linear and orthogonal expandability. It would be desirable to overcome such limitations while retaining the linear and orthogonal expansion characteristics of the known art.

The present invention provides a system which meets these criteria.

SUMMARY OF THE INVENTION

The invention comprises a multiprocessing computer system with disk array data storage systems allowing for linear and orthogonal expansion of data storage capacity and bandwidth by means of a switching network coupled between the disk array systems and the multiple processors.

More specifically, the switching network is coupled between a plurality of CPU's and a plurality of disk array systems. The switching network provides the ability for any CPU to be directly coupled to any disk array.

By using the switching network to couple multiple CPU's to multiple disk array systems, the

computer system can be configured to optimally match the I/O bandwidth of the disk array systems to the I/O performance of the CPU's.

The details of the preferred embodiment of the present invention are set forth in the accompanying drawings and the description below. Once the details of the invention are known, numerous additional innovations and changes will become obvious to one skilled in the art.

BRIEF DESCRIPTION OF THE DRAWINGS

FIGURE 1 is a block diagram of a prior art multiprocessor system.

FIGURE 2 is a block diagram of a first embodiment of the present invention.

FIGURE 3A is a block diagram of a cross-bar switching network suitable for use in conjunction with the present invention.

FIGURE 3B is a block diagram of a multi-stage switching network suitable for use in conjunction with the present invention.

FIGURE 4 is a block diagram of a second embodiment of the present invention.

Like reference numbers and designations in the drawings refer to like elements.

DETAILED DESCRIPTION OF THE INVENTION

Throughout this description, the preferred embodiment and examples shown should be considered as exemplars, rather than as limitations on the present invention.

The problems presented by the prior art in coupling a multiprocessing computer system with a disk array system are solved by the present invention by means of a novel architecture of the type shown in FIGURE 2. As in the prior art, a plurality of CPU's 1 are coupled together by at least one interprocessor bus IPB. Each CPU 1 has at least one I/O bus 3. In addition, at least one disk array 4, comprising at least one array controller 5 and a plurality of disks 6, is provided to be coupled to the CPU's 1.

In the preferred embodiment, each disk array 4 has at least two array controllers 5 to provide redundancy. The disk arrays may be of any type (e.g., RAID 1 through 5). An example of one such array is shown in pending U.S. patent application serial no. 07/270,713, entitled Arrayed Disk Drive System and Method, and assigned to the assignee of the present invention.

The problems of the prior art are specifically overcome by providing a switching network 7 that is coupled to a plurality of the CPU's 1 by corresponding CPU I/O busses 3, and to each disk array 4. The switching network 7 provides the ability for any CPU 1 to be directly coupled to any disk array 4.

The switching network 7 may be of any suitable NxN type, capable of directly coupling any node to any other node (i.e., any CPU 1 to any disk array 4). The architecture of the switching network 7 can be, for example, an NxN cross-point switch or an NxN multi-stage switch. An example of a cross-point switch architecture is shown in FIGURE 3A, which shows a plurality of nodes 10 and a corresponding plurality of communications links 11. Each node *i* 10 is coupled via an output port N_i to one communications link 11, and to each of the communications links 11 via an input port n_i through a multiplexor 12 As is known, such couplings permit each node to transfer signals through its output port N_i to any input port n_i. The selection of signal paths can be controlled by addresses from each node, in known fashion. Multiple simultaneous couplings are possible if no conflicts in addresses occurs. For example, node #1 can be coupled to node #2 while node #4 is simultaneously coupled to node #6.

An example of a multi-stage switch architecture is shown in FIGURE 3B, which shows a plurality (2,048, by way of example only) of node output ports N_i coupled to an equal number of node input ports n_i. In the example shown, 64 Stage 1 selectors 15, each 32x63 in size, permit any one of 32 inputs N_i to be coupled to any one of 63 outputs. The outputs of each Stage 1 selector 15 are coupled to each of 63 selectors 16 comprising Stage 2. The Stage 2 selectors 16 are each 64x64 in size, which permits any one of the 64 inputs to be coupled to any one of 64 outputs. In turn, the outputs of each Stage 2 selector 16 are coupled to each of 64 selectors 17 comprising Stage 3. The Stage 3 selectors 17, each 63x32 in size, permit any one of the 63 inputs to be coupled to any one of 32 outputs n_i.

Again, as is known, such couplings permit each node to transfer signals through its output port N_i to any input port n_i (other than its own, in the example shown). For example, if it is desired to couple output port N_1 to input port n_{2048}, output port N_1 is selected as the output of selector #1 in Stage 1. That output is coupled to an input of selector #63 in Stage 2, which selects that input as its output. The output of Stage 2 is coupled to the input of selector #64 in Stage 3, which selects that input as its output. The output of Stage 3 is coupled to input port n_{2048}, as desired. Again, the selection of signal paths can be controlled by addresses from each node, and multiple simultaneous couplings are possible if no conflicts in addresses occurs.

In the preferred embodiment, the switching network 7 comprises fiber optics links for high-speed data transfers. However, wired links may be used

for lower speed implementations. Also in the preferred embodiment, the switching network 7 is fault tolerant to provide continuous operation in the event of a failure of any single component. Fault tolerance of this type is well-known in the art. Alternatively, dual switching networks 7 may be provided, coupled as shown in FIGURE 3 to multiple CPU's 1 through conventional channel adapters 8, to provide redundancy.

In any case, it is preferable for the switching network 7 to have a data transmission bandwidth approximately equal to the number of nodes (i.e., coupled CPU's) multiplied by the individual I/O channel bandwidth of the CPU's 1. For example, referring to FIGURE 2, if CPU #1 is communicating with Disk Array #0, CPU #2 can communicate with Disk Array #1 at the full speed allowed by the I/O link between the two nodes, independent of the operation of CPU #1 and Disk Array #0. This characteristic provides for linear expansion of the CPU's 1 and of the disk arrays 4.

By using a switching network 7 to couple multiple CPU's to multiple disk arrays 4, the computer system can be configured to optimally match the I/O bandwidth of the disk array systems 4 to the I/O performance of the CPU's 1. For example, if an application requires a higher rate of data to be transferred to the CPU's 1, then the system can be expanded linearly in the "y" direction by adding more data storage units 6 to a disk array 4 (up to the data transfer capacity of the I/O channel 3 coupled to that disk array, or up to the data transfer capacity of the array controller 5; thereafter, additional data storage units 6 must be added to another disk array 4, or another disk array 4 must be coupled to the switching network 7). The additional data storage units 6 increase the sum of the data transfer rates of the disk arrays 4.

On the other hand, if the data transfer capacity of the disk arrays 4 exceeds the data transfer capacity of the CPU's 1, or where an application requires a higher rate of I/O to be generated than the CPU's 1 can provide, then the system can be expanded linearly in the "x" direction by coupling more CPU's 1 to the switching network 7. The additional CPU's 1 increase the sum of the data transfer rates of the CPU's as a group.

Thus, the present invention provides a means for matching the I/O bandwidth of a disk array data storage system to the I/O performance of multiple CPU's in a multiprocessor computer system having linear and orthogonal expandability.

A number of embodiments of the present invention have been described. Nevertheless, it will be understood that various modifications may be made without departing from the spirit and scope of the invention. For example, the disk array storage units need not be of the rotating disk (magnetic or optical type, but can be any type of peripheral data storage units, such as magnetic type or semiconductor memory units. Accordingly, it is to be understood that the invention is not to be limited by the specific illustrated embodiment, but only by the scope of the appended claims.

Claims

1. A multiprocessing computer system comprising:
   a. a plurality of data processing units;
   b. at least one data storage array system;
   c. switching network means, coupled to the plurality of data processing units and to at least one data storage array system, for establishing a communications link between at least one selected data processing unit and at least one data storage array system.

2. The multiprocessing computer system of claim 1, wherein the switching network means comprises a cross-point switch.

3. The multiprocessing computer system of claim 1, wherein the switching network means comprises a multi-stage switch.

4. A linearly and orthogonally expandable multiprocessing computer system comprising:
   a. at least two data processing units intercoupled by at least one interprocessor bus, the at least one bus having sufficient capacity to be coupled to at least one additional data processing unit, each data processing unit having a respective input/output data transfer rate;
   b. at least one data storage array system including at least two data storage units, at least one data storage array system having sufficient capacity to be include at least one additional data processing unit, each data storage array unit having a respective input/output data transfer rate;
   c. switching network means, coupled to the data processing units and to at least one data storage array system, for establishing a communications link between at least one selected data processing unit and at least one data storage array system, the switching network means having an input/output data transfer rate;
   wherein the input/output data transfer rate of the switching network means at least equals the sum of input/output data transfer rates of either the data processing units or the data storage array systems, and additional data storage units may be added to at least one

data storage array system to increase the sum of the input/output data transfer rates of the data storage array systems, and additional data processing units may be added to the interprocessor bus to increase the sum of the input/output data transfer rates of the data processing units.

5. The multiprocessing computer system of claim 4, wherein additional data storage units or additional data processing units are added so that the sum of the input/output data transfer rates of the data processing units is matched to approximately equal the sum of the input/output data transfer rates of the data storage array systems.

6. The multiprocessing computer system of claim 4, wherein the switching network means comprises a cross-point switch.

7. The multiprocessing computer system of claim 4, wherein the switching network means comprises a multi-stage switch.

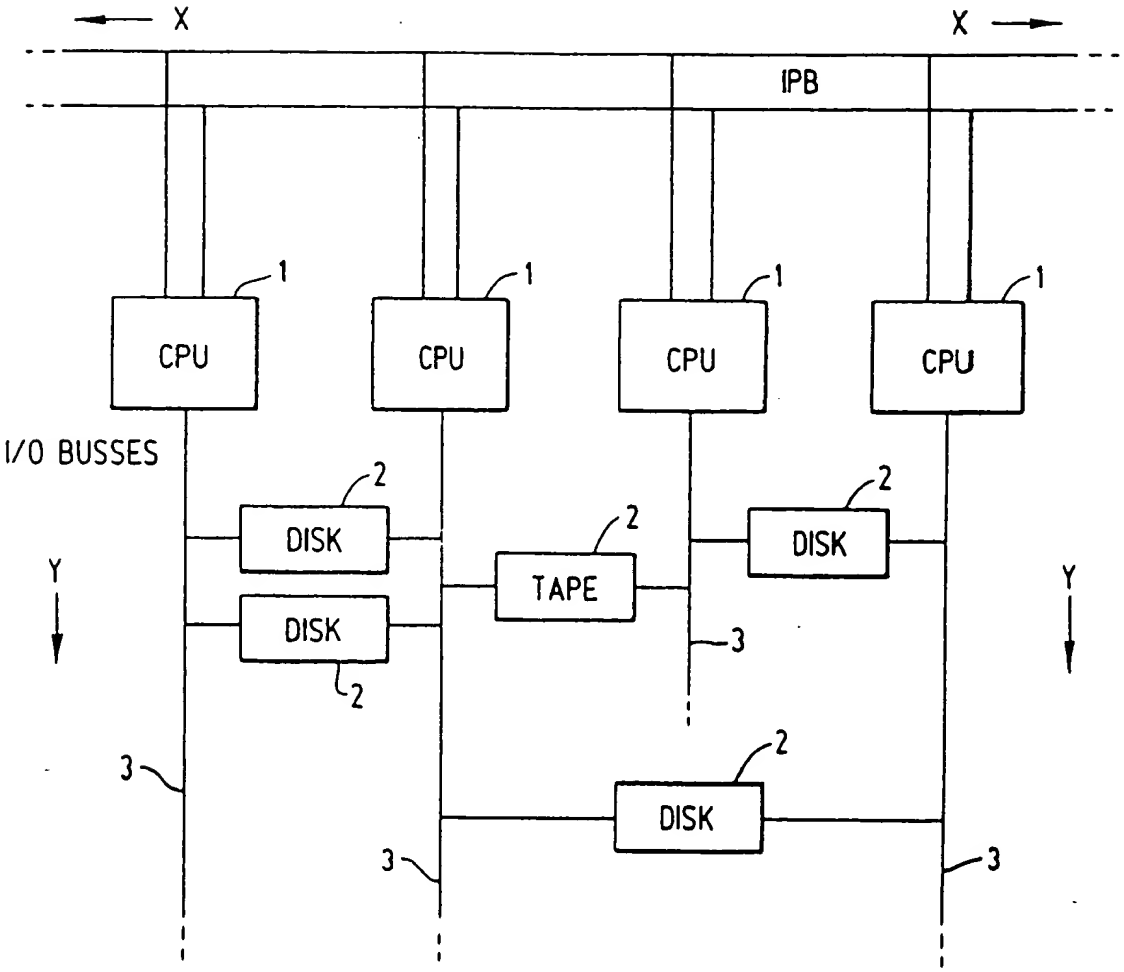

FIG. 1

PRIOR ART

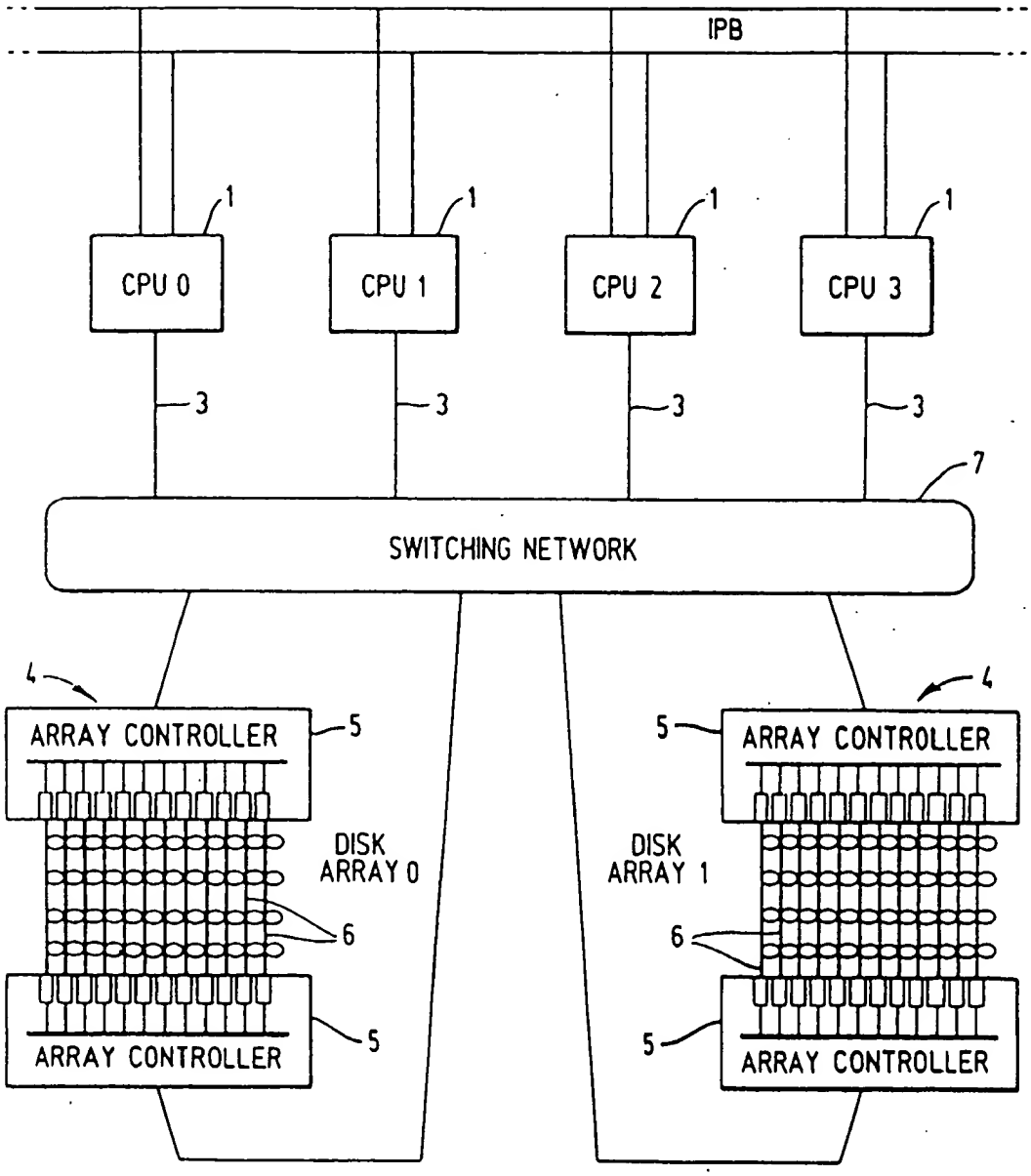
IPB
CPU 0
CPU 1
CPU 2
CPU 3
1
3
7
SWITCHING NETWORK
4
5
ARRAY CONTROLLER
DISK ARRAY 0
DISK ARRAY 1
6

FIG. 2

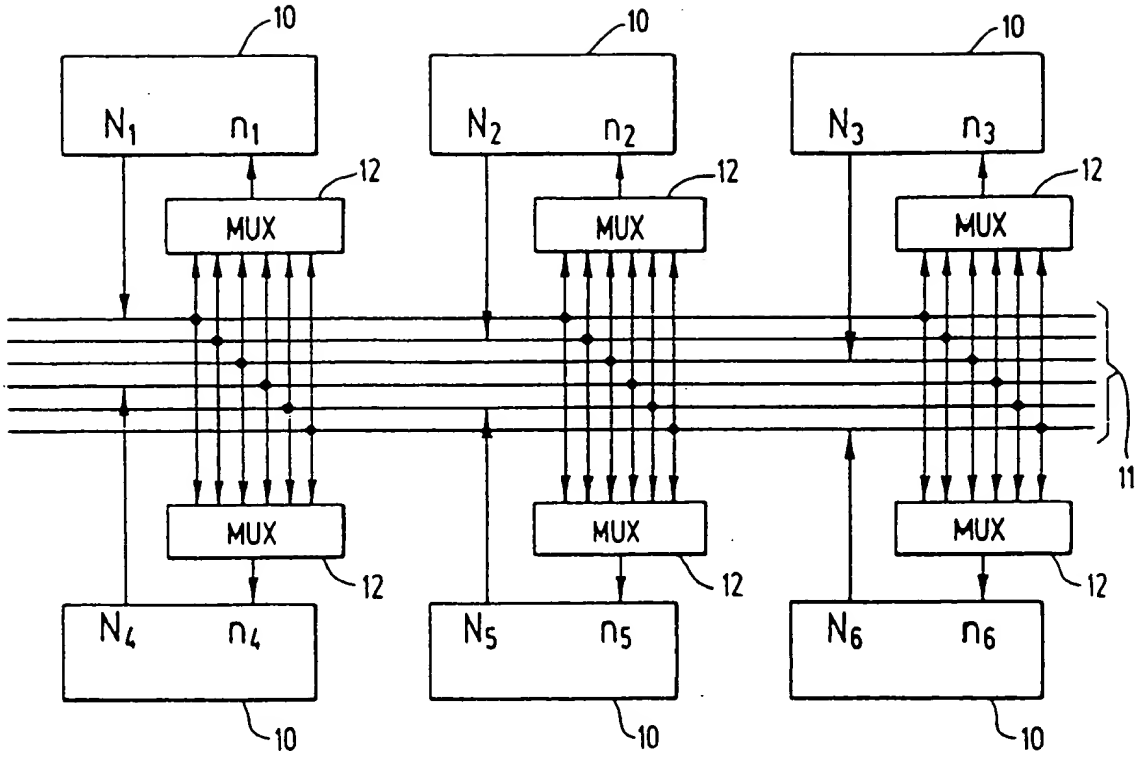
10
N1 n1
12
MUX
10
N2 n2
12
MUX
10
N3 n3
12
MUX
11
MUX
12
N4 n4
10
MUX
12
N5 n5
10
MUX
12
N6 n6
10

FIG. 3A

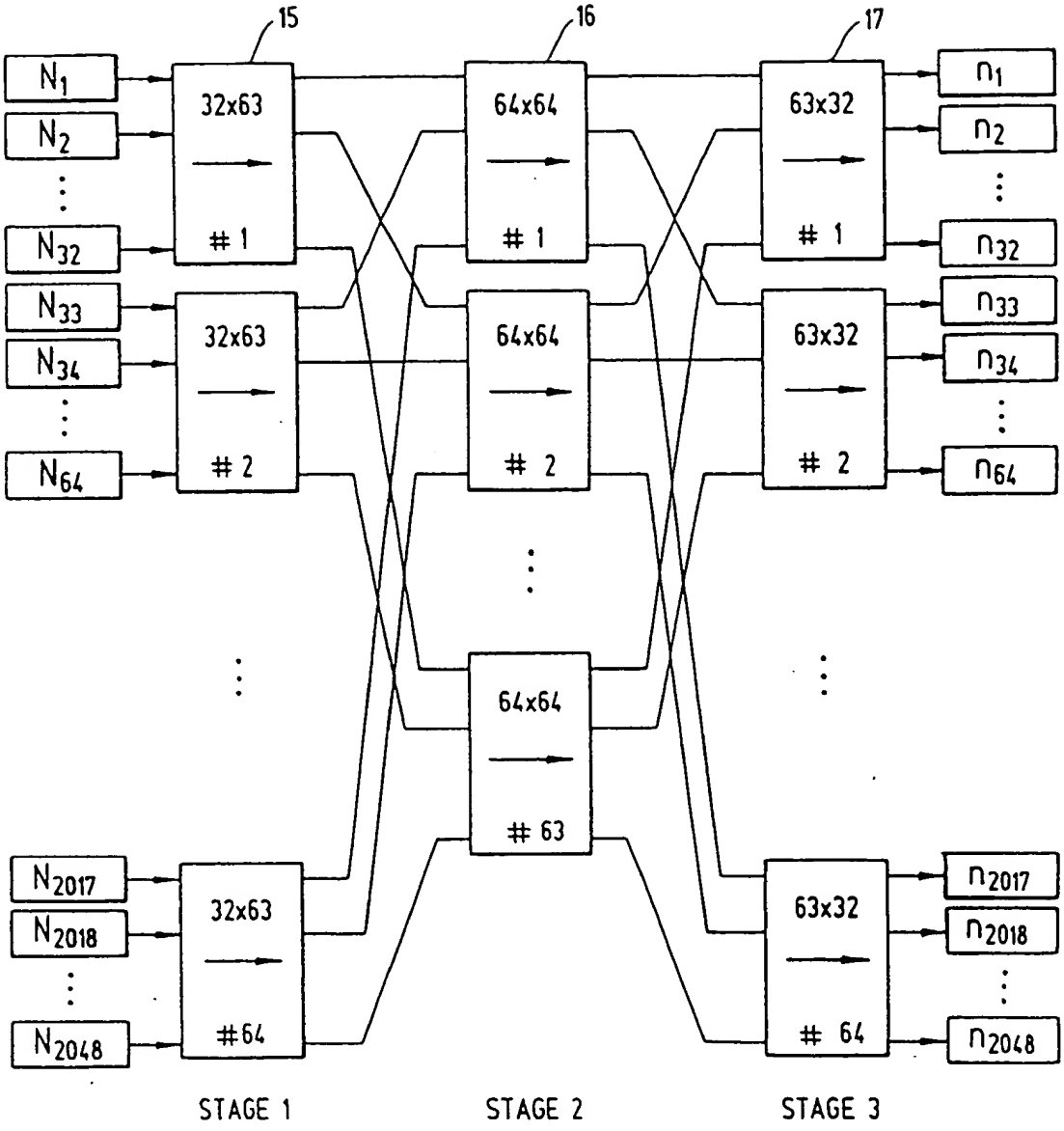
15
16
17
N1
N2
N32
N33
N34
N64
N2017
N2018
N2048
32x63
1
32x63
2
32x63
#64
64x64
1
64x64
2
64x64
63
63x32
1
63x32
2
63x32
64
n1
n2
n32
n33
n34
n64
n2017
n2018
n2048
STAGE 1
STAGE 2
STAGE 3

FIG 3B

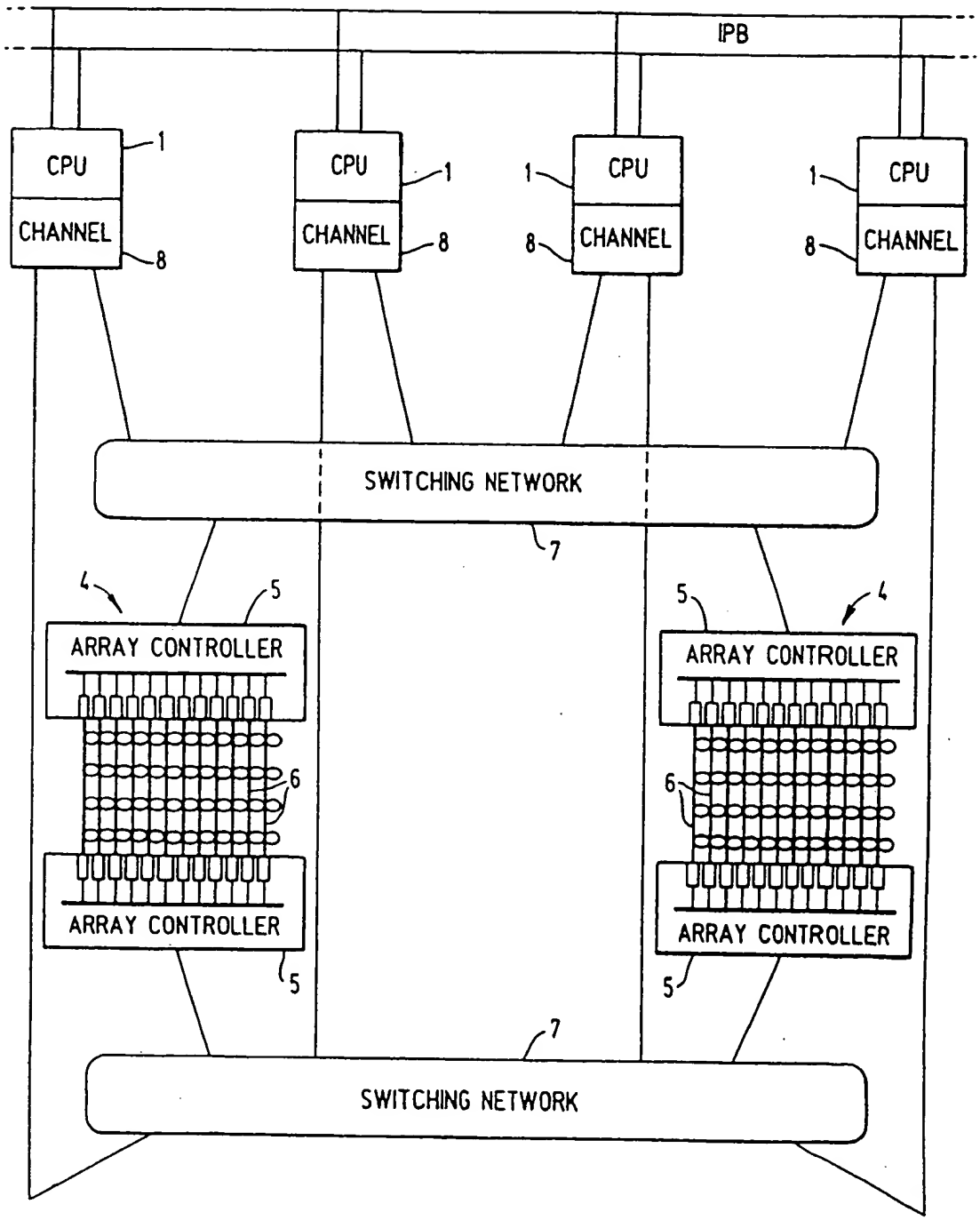
IPB
CPU
CHANNEL
1
8
SWITCHING NETWORK
7
4
5
ARRAY CONTROLLER
6

FIG. 4

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Publication number: 0 535 807 A3

(12) EUROPEAN PATENT APPLICATION

(21) Application number: 92308079.0

(51) Int. Cl.5: G06F 15/16

(22) Date of filing: 07.09.92

(30) Priority: 01.10.91 US 769538

(43) Date of publication of application:
07.04.93 Bulletin 93/14

(84) Designated Contracting States:
DE FR GB IT

(88) Date of deferred publication of the search report:
22.09.93 Bulletin 93/38

(71) Applicant: TANDEM COMPUTERS INCORPORATED
10435 N. Tantau Avenue
Cupertino, California 95014-0709(US)

(72) Inventor: Walker, Mark
20000 Gist Road
Los Gatos, California 95030(US)
Inventor: Lui, Albert S.
3164 Heritage Valley Drive
San Jose, California 95148(US)
Inventor: Sammer, Harald
Am Eschenhorst 10
W-6382 Friedrichsdorf(DE)
Inventor: Chan, Wing M.
7922 Kemper Court
Pleasanton, California 94588(US)
Inventor: Fuller, William T.
1536 Shasta Avenue
San Jose, California 95126(US)

(74) Representative: Allman, Peter John et al
MARKS & CLERK Suite 301 Sunlight House
Quay Street
Manchester M3 3JY (GB)

(54) Linear and orthogonal expansion of array storage in multiprocessor computing systems.

(57) A multiprocessing computer system with data storage array systems allowing for linear and orthogonal expansion of data storage capacity and bandwidth by means of a switching network coupled between the data storage array systems and the multiple processors. The switching network provides the ability for any CPU to be directly coupled to any data storage array. By using the switching network to couple multiple CPU's to multiple data storage array systems, the computer system can be configured to optimally match the I/O bandwidth of the data storage array systems to the I/O performance of the CPU's.

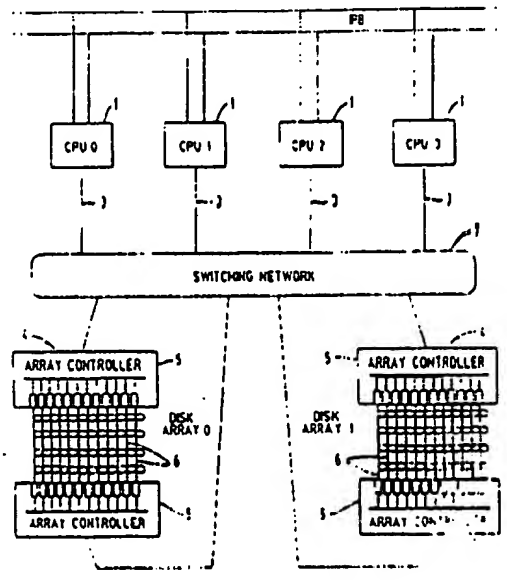

FIG. 2

EP 0 535 807 A3

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 92 30 8079

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int. Cl.5)
X Y	WO-A-9 114 229 (SF2 CORPORATION) * the whole document *	1-3 4-7	G06F15/16 G06F13/40
Y	EP-A-0 380 851 (DIGITAL EQUIPMENT CORPORATION) * column 1, line 1 - column 3, line 15 *	4-7	
A,D	US-A-4 228 496 (KATZMAN) * column 3, line 27 - line 38; figure 1 * * column 6, line 55 - column 7, line 57 *	4-7	
A	WO-A-9 113 399 (SF2 CORPORATION) * the whole document *	1-7	
A	PROCEEDINGS OF THE 16TH VLDB CONFERENCE 1990, BRISBANE, AUSTRALIA pages 148 - 161 J. GRAY ET AL. 'Parity Striping of Disc Arrays: Low-Cost Reliable Storage with Acceptable Throughput'		

TECHNICAL FIELDS SEARCHED (Int. Cl.5)
G06F

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	28 JULY 1993	ABSALOM R.

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P0401)